Journal of Bioinformatics and Computational Biology
Vol. 5, No. 5 (2007) 1069–1085

ICP Imperial College Press
www.icpress.co.uk

MULTIMODAL PHYLOGENY FOR TAXONOMY: INTEGRATING INFORMATION FROM NUCLEOTIDE AND AMINO ACID SEQUENCES

MANUELE BICEGO

Dip. di Economia Impresa e Regolamentazione, University of Sassari
via Torre Tonda, 34, 07100 Sassari, Italy
bicego@uniss.it

FRANCO DELLAGLIO* and GIOVANNA E. FELIS†

Dip. Scientifico e Tecnologico, University of Verona
Strada Le Grazie, 15, 37135 Verona, Italy
**dellaglio@sci.univr.it*
†*felis@sci.univr.it*

Received 3 October 2006
Revised 2 July 2007
Accepted 8 July 2007

The crucial role played by the analysis of microbial diversity in biotechnology-based innovations has increased the interest in the microbial taxonomy research area. Phylogenetic sequence analyses have contributed significantly to the advances in this field, also in the view of the large amount of sequence data collected in recent years. Phylogenetic analyses could be realized on the basis of protein-encoding nucleotide sequences or encoded amino acid molecules: these two mechanisms present different peculiarities, still starting from two alternative representations of the same information. This complementarity could be exploited to achieve a multimodal phylogenetic scheme that is able to integrate gene and protein information in order to realize a single final tree. This aspect has been poorly addressed in the literature. In this paper, we propose to integrate the two phylogenetic analyses using basic schemes derived from the multimodality fusion theory (or multiclassifier systems theory), a well-founded and rigorous branch for which its powerfulness has already been demonstrated in other pattern recognition contexts. The proposed approach could be applied to distance matrix–based phylogenetic techniques (like neighbor joining), resulting in a smart and fast method. The proposed methodology has been tested in a real case involving sequences of some species of lactic acid bacteria. With this dataset, both nucleotide sequence– and amino acid sequence–based phylogenetic analyses present some drawbacks, which are overcome with the multimodal analysis.

Keywords: Phylogeny; sequence analysis; fusion approaches; pattern recognition.

1. Introduction

Microbial diversity is one of the major sources for the discovery and exploitation of novel biotechnological innovations.[1] For this reason, the cataloging of that diversity

is equally fundamental. The analysis of diversity, i.e. the delineation of taxa and their arrangement in an ordered scheme, is referred to as taxonomy or systematics. In the last 20 years, a major advance in microbial taxonomy arose from the analyses of the molecular elements of the cell, and from phylogenetic analyses in particular. Since the first proposal of using molecular sequences as phylogenetic markers, different molecules have been targeted: rRNA genes, polypeptides, protein-encoding genes, noncoding sequences, etc.[2–5]

The enormous amount of gene and genome sequence data available to date, due to the lowered costs of the DNA sequencing techniques, together with the easier and more objective comparison of taxa based on sequence data[6] have made it possible to reanalyze the phylogenetic and taxonomic relationship of taxa at all taxonomic levels, also outside the microbial context.[7] In a taxonomic perspective, the important points of phylogenetic sequence analysis are (1) the reliability of the obtained groupings, and (2) the clear resolution of the terminal nodes of the tree. Both aspects are related either to the kind of molecule analyzed (16S rRNA, protein-encoding genes, noncoding sequences) or to the way the phylogenetic analysis is performed. Actually, this analysis could be based on the calculation of a matrix from the data, representing the distance between each pair of aligned sequences, and subsequently transforming the matrix into a tree. Various distance measures have been proposed to this aim, which are based on different models of nucleotide substitution or amino acid replacement (e.g. Refs. 8–15). Alternatively, the phylogenetic tree could be derived by directly finding the tree topology best fitting the data; this category of methods groups parsimony and maximum likelihood.[16]

The kind of molecule chosen for the analysis is crucial for the extent of information that could be extracted from it: amino acid sequences are useful to depict phylogenetic relationships at higher taxonomic ranks or levels, but they often overestimate the relatedness of closely related sequences since synonymous mutations (i.e. DNA substitutions not leading to a mutation in the amino acid sequence) are not recorded; this often does not allow resolution of terminal branches. On the nucleotide side, protein-encoding genes tolerate synonymous substitutions, linked to the different codon usage in different organisms: this allows the differentiation of closely related taxa (e.g. Ref. 17). This result is often not pursued with the analysis of 16S rRNA gene sequences, which encode for functional ribosomal RNA and are very slowly diverging sequences, with an informative content similar to that of protein sequences. However, even if protein-encoding DNA sequence analysis allows fine resolution of closely related taxa, they could produce inconsistent groupings if sequence characteristics such as guanine-cytosine (GC) content are not properly addressed.

Summarizing and considering protein-encoding genes, the DNA, and the amino acid sequences are two sides of the same coin, connected by the translation mask, representing two different representations of the same information. In practical cases, the choice of using the amino acid or the nucleotide phylogeny is typically related to the context: the most used solution is to realize both, manually trying to

derive an *a posteriori* final agreement. Sometimes, the use of models for nucleotide sequence analysis (e.g. Ref. 13, which accounts for compositional bias of sequences) permits the exploitation of information on the encoded amino acid sequences, albeit only implicitly.

For all of these reasons, the realization of an automatic and explicit way of integrating protein and gene phylogenies could be of great impact in this research field, and is the aim of the present paper. This aspect has been poorly addressed in the literature, and typical approaches implemented in current softwares are mostly based on heuristic rules, lacking a rigorous and well-founded theory. An exception is represented by the MrBayes3 software,[18] based on the rigorous Bayesian theory, which allows the combination of different kind of data. Nevertheless, this approach presents some disadvantages, especially from a practical and operative point of view: the first concern is relative to the several additional parameters to be set (like the prior probabilities), most of which are crucial. Moreover, the derived tree is the result of a complex optimization process (based on Markov chain Monte Carlo), which in some cases does not converge to an acceptable solution. Finally, it is very slow and thus inapplicable for large datasets. For all of these reasons, taxonomists typically resort to distance matrix–based methods.

In this paper, a novel method for integrating gene and protein information in phylogenetic analyses is proposed, which has different appealing characteristics: (1) it is based on a rigorous and well-founded theory, namely, the fusion theory[a]; (2) one of the proposed methods clearly derives from a biological observation; (3) it is a distance matrix–based approach, so it is fast and easy to apply; and (4) the result is a single final phylogenetic tree, obtained from both nucleotide and amino acid analyses by the application of fusion strategies well known in the field of computer science, but poorly applied in biologically relevant contexts. Fusion theory aims at integrating the possibly complementary information provided by different methodologies in a particular problem, exploiting the different peculiarities of the fused techniques. This theory, first introduced in the classification context[19–21] and recently also in the clustering context (Refs. 22 and 23 as well as references therein), seems to be particularly suited for the context we are investigating. In particular, the information fusion could be performed at three different levels[24]: data or feature level, where feature representations are combined; score level, where scores derived from different modalities (e.g. similarities) are composed to get a new score; and decision level, where the final outputs of multiple strategies are consolidated.

The novel multimodal fusion approach proposed in this paper belongs to the second category, and is aimed at integrating gene- and protein-derived similarity matrices, resulting in a score level fusion. In general, fusion at score level is preferred,[25,26] since it is relatively easy to access and combines scores produced by the different modalities; moreover, in some studies, its superiority against feature-level fusion and decision-level fusion has been reported (e.g. Ref. 27). A score-level

[a]Also referred, in other contexts, as multimodality approach or multiple classifier system.

fusion technique has been very recently proposed in the context of phylogenomics,[28] which is nevertheless an extension of phylogeny. This scenario presents several different problems such as the lack of data in datasets and the problematic combination of genes with different histories (resulting in different mutation rates). In our approach, we do not combine different molecules, but different representations of the same molecule. The aim is therefore to extract as much phylogenetic signal as possible from a single trait, circumventing the drawbacks due to the type of molecule analyzed and to the different methods of analysis. The intervention we propose in the present study is therefore a step upstream to the combination of results of different genes, i.e. phylogenomics.

In this study, different score-level fusion strategies were tested in a real case involving a set of sequences of lactic acid bacteria species, with the peculiarity of an unequal base composition. In this case, it is known that the majority of distance formulas have some limits. It has been shown that the fused tree overcomes these limits and allows, at the same time, the recognition of peculiarities of individual taxa. In other words, it is able to maintain the appealing characteristics of both nucleotide and amino acid trees, recovering from the single-modality drawbacks.

2. Multimodal Fusion

Fusion theory starts from the following rationale: different methodologies could be designed and tested to solve a practical problem (classification or clustering). Although one of these methodologies would yield the best performance, the sets of patterns wrongly treated by different methodologies would not necessarily overlap. This suggests that different methodologies potentially offer complementary information, and an approach which fuses decisions taken by different sources could lead to better performance. Multimodality approaches have been successfully employed in different pattern recognition fields: two significant examples are the audio-visual joint scene analysis (see, for example, Refs. 29–32), and the multimodal biometrics (see, for example, Refs. 33–35).

As stated in Sec. 1, fusion could be performed at feature level, score level, or decision level. Here, we concentrate on fusion at the score level, i.e. on the combination of similarity (or dissimilarity) measures originating from different modalities. In this context, two scenarios are possible[20]: (1) all of the methodologies (classifiers) use the same representation for the input data; or (2) each methodology has its own one, i.e. the measurements extracted from each object are unique to each methodology. The fusion schemes adopted in this paper are related to this second scenario, since the nucleotide and amino acid phylogenies are based on different characterizations of the same taxon.

In the context of fusion, different basic rules have been proposed and tested. In this paper, we investigated five different strategies. Four of them are standard rules, directly derived from the fusion theory.[20] They are typically well performing in different applications, but they lack a clear biological meaning. For this reason,

we introduced a fifth one, clearly understandable from a biological standpoint. All of these techniques take as input the two similarity matrices derived from both the nucleotide sequence– and the amino acid sequence–based phylogenies, which we called D_{NT} and D_{AA}, respectively; and produce as output a new single similarity matrix, denoted as D_{Fusion}.

2.1. *Standard rules*

We applied the following four standard rules:

(1) **Mean rule**: For each pair of sequence, the new distance is the mean of the nucleotide and amino acid ones ($1 \leq i, j \leq N$):

$$D_{\mathrm{Fusion}}(i,j) = \frac{1}{2}(D_{\mathrm{NT}}(i,j) + D_{\mathrm{AA}}(i,j)), \tag{1}$$

where N is the number of sequences. Actually, the mean rule is equivalent to the sum one,[20] since the tree does not change in a qualitative way (the distance only scaled to a factor of 2); nevertheless, the mean rule maintains the distance measures in the same range of the input ones.

(2) **Prod rule**: For each pair of sequence, the new distance is the product of the input ones:

$$D_{\mathrm{Fusion}}(i,j) = D_{\mathrm{NT}}(i,j) \times D_{\mathrm{AA}}(i,j). \tag{2}$$

(3) **Max rule**: For each pair of sequence, the new distance is the maximum between the nucleotide distance and the amino acid one:

$$D_{\mathrm{Fusion}}(i,j) = \max(D_{\mathrm{NT}}(i,j), D_{\mathrm{AA}}(i,j)). \tag{3}$$

(4) **Min rule**: For each pair of sequence, the new distance is the minimum between the two input distances:

$$D_{\mathrm{Fusion}}(i,j) = \min(D_{\mathrm{NT}}(i,j), D_{\mathrm{AA}}(i,j)). \tag{4}$$

A common theoretical framework justifying these rules (and others) can be found in Ref. 20. A recent analysis of the mean rule (and in general of the linear combiners for multiple classifier systems) can be found in Ref. 36. Even if more complex rules more explicitly related to clustering applications have been recently proposed (see, for example, Refs. 22 and 23), these simple rules typically permit us to obtain very satisfactory results.[34]

2.2. *An adaptive fusion rule*

The mean rule could also be considered as the basic version of the so-called linear combiners, which assign a different weight to each methodology.[36] This rule, in our context, becomes

$$D_{\mathrm{Fusion}}(i,j) = (1-\alpha)D_{\mathrm{NT}}(i,j) + \alpha D_{\mathrm{AA}}(i,j), \tag{5}$$

with $0 \leq \alpha \leq 1$.

The main problem in this case is to choose the right value of α. One solution is to set α with a fixed value (as done in Ref. 28, related to the reliability of the protein-based analysis, where $1 - \alpha$ is the counterpart for the nucleotide-based analysis. Here, we adopt a different strategy, based on a well-known fact: the amino acid analysis is more reliable for very distant taxa, whereas the nucleotide one is more discriminative for related taxa. Following this rationale, we define a different weight for each pair of analyzed taxa, giving more emphasis to the amino acid part if the taxa are unrelated or to the nucleotide one if the taxa are strictly related. The rule therefore becomes

$$D_{\text{Fusion}}(i,j) = (1 - \alpha_{ij}) D_{\text{NT}}(i,j) + \alpha_{ij} D_{\text{AA}}(i,j), \tag{6}$$

with $0 \leq \alpha_{ij} \leq 1$.

There are several possible options for defining α_{ij} that reflects this fact. The first intuitive solution is

$$\alpha_{ij} = \frac{R(i,j)}{\max_{h,k} R(h,k)}, \tag{7}$$

where $R(i,j)$ is the relatedness of the taxa i and j. In our case, we define the relatedness as the amino acid distance, i.e.

$$R(i,j) = D_{\text{AA}}(i,j).$$

The solution for Eq. (7) has two drawbacks. First, if in the dataset there is an outgroup, there will be few large distances and many small distances, and all α_{ij} will be small. This could be avoided by linking the value of $\alpha_{ij} = 0.5$ to the median of the distances (the median is a robust estimation of the mean). Second, since the protein is more reliable than the gene, we should guarantee to always fuse a minimal amount of information from the amino acid analysis. This could be guaranteed by imposing

$$I_{\min} \leq \alpha_{ij} \leq 1,$$

where $0 \leq I_{\min} \leq 1$ is the minimum amount of information assured to the amino acid analysis. The final formulation is therefore

$$\alpha_{ij} = I_{\min} + \left(\frac{R(i,j) * 0.5}{\text{median}_{h,k} R(h,k)} \right) (1 - I_{\min}). \tag{8}$$

In this way, we used protein-based analysis to depict the general behavior of the tree, leaving to the gene-based one the resolution of very close taxa.

2.3. *Score normalization*

A crucial issue to be solved in the context of fusion at the score level is the normalization problem,[24] which regards the transformation of the scores of different modalities in a common range prior to combining them (for example, if one score has values in the range [0,1] and another in the range [0,100000], the fusion will be completely driven by the second).

3. Experimental Results

The proposed approach was tested in a real case involving a set of sequences from some species of lactic acid bacteria. In particular, the dataset was composed by 71 partial *recA* gene sequences (summarized in Table 1) of 564 bp and, consequently, 188 amino acid positions.

Sequences were aligned with the ClustalX[37] program and gap-containing columns were removed. Alignment did not introduce gaps; therefore, there were no consequences on the translation frame of the nucleotide sequences. The phylogenetic analyses were performed with Kimura, Logdet, Jukes–Cantor, and F84 models for nucleotide sequences; JTT, PMB, PAM, and Kimura for amino acid sequences;

Table 1. Dataset used in the experimental session.

Label name	GenBank acc. no.	Label name	GenBank acc. no.
Lacidipisc	AJ621616	Lmali	AJ621655
Lagilis	AJ621617	Lmaltaromi	AJ621690
Lamylophil	AJ621621	Lmucosae	AJ621657
Lamylovoru	AJ621622	Lmurinus	AJ621658
Lanimalis	AJ621623	Lpartolerans	AJ621663
Laviarar	AJ621624	Lparac	AJ621664
Lbrevis	AJ621625	Lparabuchn	AJ621661
Lbuchneri	AJ621626	Lparaplant	AJ621662
Lcasei334	AJ621627	Lpentosus	AJ621666
Lcatenafor	AJ621629	Lperolens	AJ621667
Lcellobios	AJ579535	Lplantarum	AJ621668
Lcolehomin	AJ621630	Lpontis	AJ621669
Lcollinoid	AJ621631	Lpsittaci	AJ621670
Lcrispatus	AJ621632	Lreuteri	AJ621672
Lcurvcurv	AJ621633	Lruminis	AJ621673
Lcypricase	AJ621634	Lsharpeae	AJ621675
Ldelbdelb	AJ586863	Lsuebicus	AJ621676
Ldellactis	AJ586865	Lvaccinost	AJ621678
Ldelbulg	AJ586864	Lvaginalis	AJ621679
Ldiolivora	AJ621635	Pdamnosus	AJ621694
Ldurianis	AJ621636	Pinopinatu	AJ621697
Lfarcimini	AJ621638	Pparvulus	AJ621698
Lfermentum	AJ579534	Ppentosace	AJ621699
Lfornicali	AJ621639	Purinaequi	AJ621700
Lfrumenti	AJ621640	Lncarnosum	AJ621682
Lfuchuensi	AJ621641	Lncitreum	AJ621688
Lgallinaru	AJ621642	Lnfallax	AJ621684
Lgasseri	AJ621643	Lnmesdex	AJ621685
Lgraminis	AJ621644	Lnmescrem	AJ621687
Lhamsteri	AJ621646	Lnmesmes	AJ621686
Lhelveticu	AJ621645	Lnpseudome	AJ621683
Lhilgardii	AJ621647	Ooeni	AJ621689
Lintestina	AJ621654	Wkandleri	AJ621692
Ljensenii	AJ621648	Wminor	AJ621693
Lkefiri	AJ621650	Cdivergens	AJ621691
Lkunkeei	AJ621652		

and neighbor joining as the clustering method, as implemented in Phylip Version 3.6.[b] The fused distance matrices were computed with the Matlab code. All distance formulas tested produced almost the same tree in terms of topology and group composition. As an example, the gene sequence–derived dendrogram obtained with the LogDet model is shown in Fig. 1, whereas the amino acid–derived one (PMB model) is presented in Fig. 2.

In order to highlight the drawbacks of the single-modality phylogenetic trees, consider sequences labeled as Lsharpeae, Ldelbdelb, Ldelbulg, Ldellactis, Lpontis, Lfermentum, and Lcellobios (group 1 in Fig. 1): the group does not agree with clusters obtained with other genes such as 16S rRNA and with phenotypic traits of the taxa.[38] Moreover, this group is split into four subgroups (1a, 1b, 1c, and 1d) in the amino acid tree (Fig. 2), consistently with the above-mentioned data. The nucleotide sequences are characterized by a similar GC content and accordingly similar codon usage and nucleotide frequencies, which bias the analysis of the nucleotide sequences; however, this is overcome by the analysis of the amino acid sequences.

On the other side, species clustered in group 2 (*Lactobacillus pentosus*, *Lactobacillus plantarum*, *Lactobacillus paraplantarum*) are clearly separated only in the gene sequence–derived tree: those species are closely related and mutations in the gene sequences are mainly synonymous substitutions.[17] Therefore, distances are not determinable in the amino acid–derived tree. From these observations, it derives that the two single-modality phylogenetic trees made some approximations in two different parts of the tree. The complementarity of these errors is the correct premise for applying the fusion strategy.

The multimodal trees were obtained using the five rules described in Sec. 2 to combine the scores of the nucleotide and amino acid analyses, for different models. The scores were normalized using their averages. Figures 3 and 4 show the multimodal tree obtained by fusing the LogDet and the PBM distance matrices with the mean rule and the adaptive rule, as described in Eq. (1) and Eq. (6), respectively; they produced the best results. Actually, in the reported trees, the spreading of group 1 throughout the tree is maintained; therefore, the biases of the DNA analysis seem to be overcome. Considering group 2, the species are clearly separated, bypassing the low resolution of the amino acid sequence–based tree. Concerning branch lengths and group composition, other rules produced unsatisfactory results. In particular, the Min and the Prod rule results were driven by the amino acid analysis, showing a correct topology but a poor resolution. This behavior is expected, since both rules are mostly influenced by the lowest scores (amino acid ones, due to the degeneration of the genetic code). Vice versa, the Max rule result was led by the nucleotide analysis, hence maintaining incorrect grouping but clearer resolution of branches.

[b]All information on software and models can be found at http://evolution.gs.washington.edu/phylip.html.

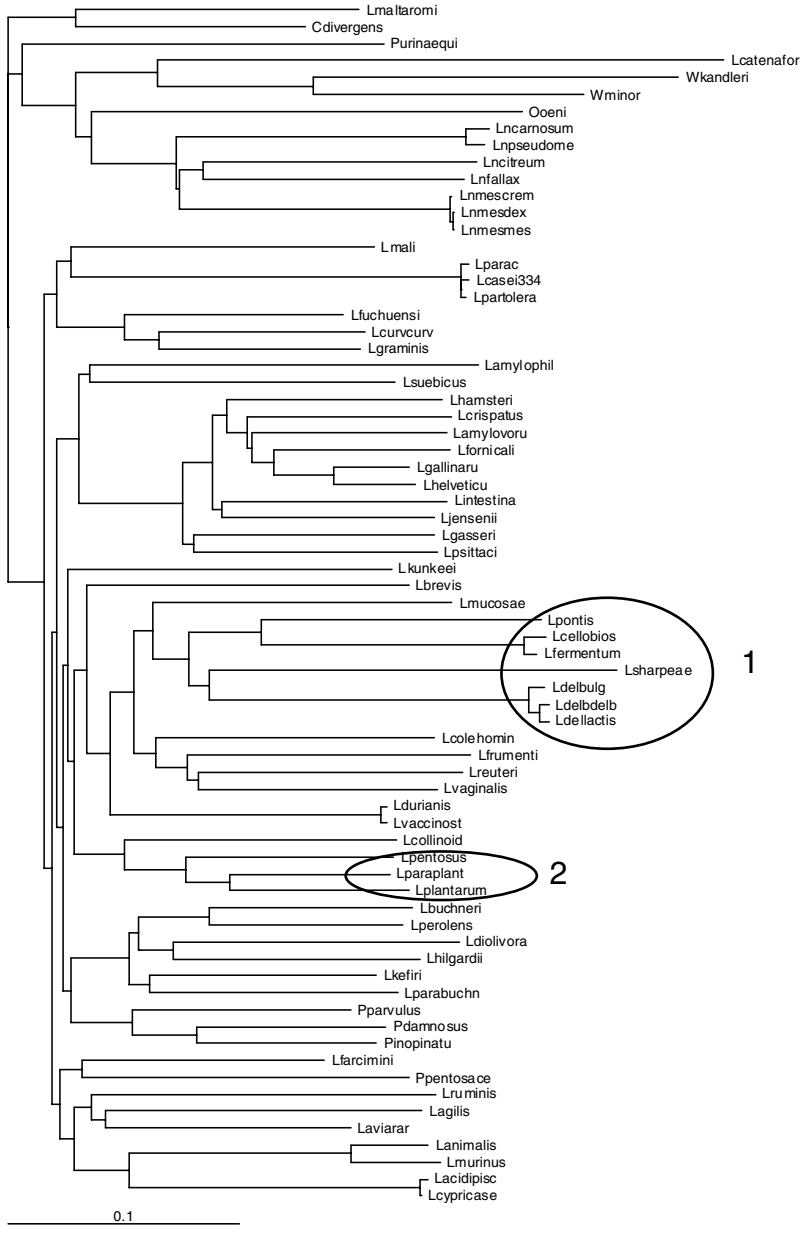

Fig. 1. Dendrogram derived from the nucleotide sequence analysis. Two groups are highlighted. The first includes the sequences of *Lactobacillus sharpeae* (labelled Lsharpeae), *Lactobacillus delbrueckii* subsp. *delbrueckii* (Ldelbdelb), *Lactobacillus delbrueckii* subsp. *bulgaricus* (Ldelbulg), *Lactobacillus delbrueckii* subsp. *lactis* (Ldellactis), *Lactobacillus pontis* (Lpontis), *Lactobacillus fermentum* (Lfermentum), and *Lactobacillus cellobiosus* (Lcellobios). The second comprises *Lactobacillus pentosus* (Lpentosus), *Lactobacillus plantarum* (Lplantarum), and *Lactobacillus paraplantarum* (Lparaplant).

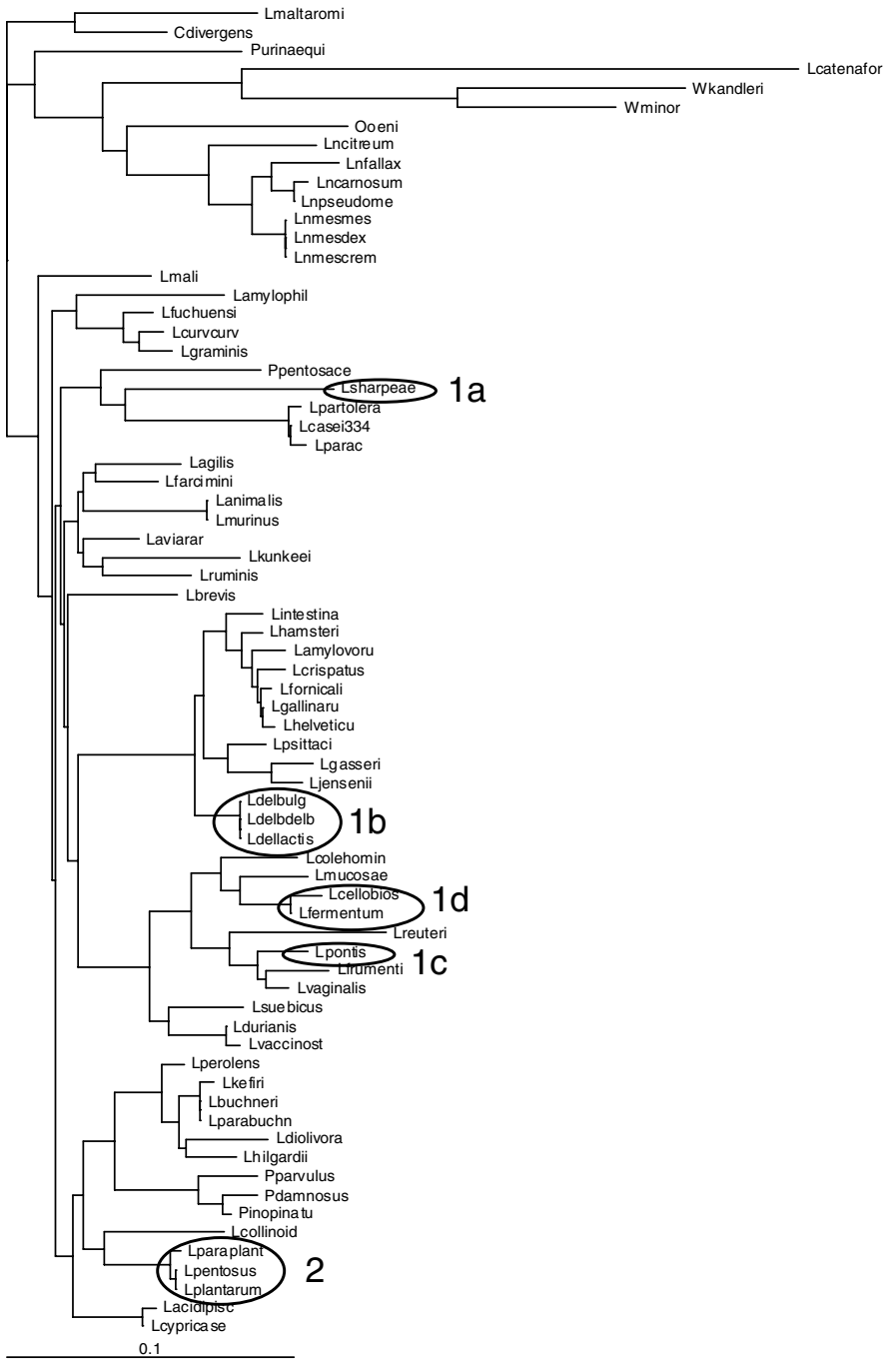

Fig. 2. Dendrogram obtained from the amino acid sequence analysis. Note that group 1 of Fig. 1 is split into four different parts of the tree (subgroups 1a, 1b, 1c, and 1d). Group 2 remains unaltered, but distances are lost.

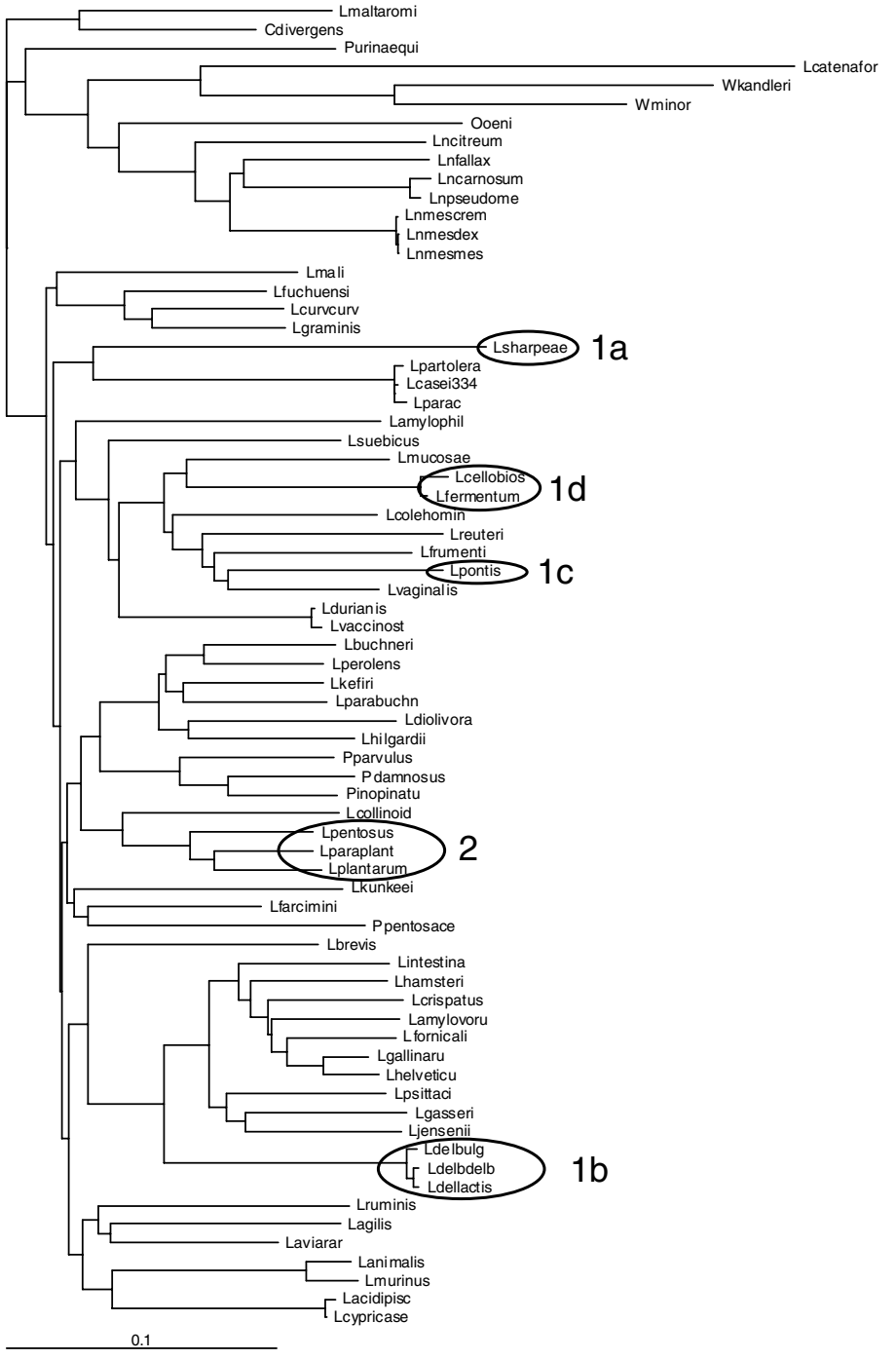

Fig. 3. Dendrogram obtained from the multimodal analysis (mean rule): the split of the group 1 is maintained, and the distances in group 2 are regained.

Fig. 4. Dendrogram obtained from the multimodal analysis (adaptive rule, with $I_{\min} = 0.3$): the split of the group 1 is maintained, and the distances in group 2 are regained.

Looking more in detail at the nucleotide-based trees and multimodal trees, some indistinguishable leaves are still present: Lnmesmes, Lnmesdex, and Lnmescrem are sequences belonging to different strains of the same species (*Leuconostoc mesenteroides*), and their high similarity is expected. The same situation characterizes Ldelbulg, Ldellactis, and Ldelbdeld (*Lactobacillus delbrueckii*) as well as Lparac, Lpartolerans, and Lcasei334 (*Lactobacillus paracasei*). Other unresolved groups contain taxa that have been recently reclassified as belonging to the same species include the following: *Lactobacillus fermentum* (Lfermentum) and *Lactobacillus cellobiosus* (Lcellobios),[39] *Lactobacillus cypricasei* (Lcypricase) and *Lactobacillus acidipiscis* (Lacidipisc),[40] and *Lactobacillus vaccinostercus* (Lvaccinost) and *Lactobacillus durianis* (Ldurianis).[41]

4. Discussion

In this preliminary study of the application of fusion techniques to phylogenetic analysis, we investigated a few models and applied a distance-based clustering technique in order to analyze as clearly as possible the effects of score-level fusion. The results of the analysis are encouraging, since single-modality errors are overcome in the multimodal tree. This is possible thanks to the complementary behavior of the single modalities, which show approximations in different parts of the tree. Obviously, this complementarity is crucial in order to make the fusion strategy advantageously working; if not present, no gain could be obtained with the fusion.

Even if the analysis is somewhat preliminary, some precise considerations could be drawn at this stage. The first concerns the analysis of the multimodal tree: by comparing it with the amino acid sequence–based analysis one, we could observe that there is a perfect agreement in terms of group composition. Although the order of ramifications is not completely preserved, no conclusions can be drawn without a deeper analysis on the significance and on the robustness of the ramifications (for example, with bootstrap). Moreover, as stated before, it is evident that the species in group 2 are clearly separated in the multimodal tree, bypassing the low resolution of the amino acid sequence–based tree. On the other hand, the presence of some unresolved leaves allows the immediate recognition of indistinguishable taxa.

Another consideration regards the nucleotide sequence–based phylogenetic analysis: we are aware that more sophisticated models could effectively deal with a dataset with unequal base composition, such as the Galtier–Gouy model proposed in Ref. 13. Nevertheless, the simplicity of the present analysis permits us to clearly evidence the actual gain obtained by the fusion strategy. Even if the multimodal analysis is based on a nonoptimal nucleotide analysis (based on an inappropriate model), the obtained result outperforms single-modality outcomes.

A further consideration concerns the fusion strategies: in our experiment, different fusion techniques did not always provide the same results, as expected. In particular, we observed that the best result was obtained with the mean rule (sum) and with the averaged mean rule (the adaptive one). The introduction of the adaptive

fusion rule did not significantly improve the results of the multimodal analysis with respect to the sum. Even if very simple, the sum scheme permits us to obtain very satisfactory results. On the other hand, the worst result has been obtained with the Prod rule: this represents a very restrictive rule, since low similarities in one methodology drive the whole fusion result to low values. These outcomes are in line with results obtained in other contexts (e.g. Refs. 20 and 34).

As a final comment, we are aware that with the proposed methodology we are employing the data sequence twice, combining the results. Data reuse is surely an open issue in the multiclassifier systems (see, for example, the discussion in Ref. 42), even if in the specific context the problems could be avoided with a careful estimation of the parameters.

5. Conclusions

In this paper, a novel multimodal phylogeny scheme has been proposed, aimed at fusing amino acid and nucleotide information at the score level. The proposed approach is based on a well-founded theory and permits the integration of distance matrix–based phylogenetic schemes, thus resulting in a fast and intuitive method. Different basic fusion schemes have been analyzed and tested in a real case involving sequences of some lactic acid bacteria species. In this dataset, both nucleotide and amino acid phylogenetic analyses have some drawbacks, which are recovered by the multimodal analysis.

The obtained preliminary result is very promising, and encourages us to go further in this direction. There are several issues to be investigated, of both practical and theoretical types. From a practical point of view, it could be interesting to statistically validate the groups obtained by the multimodal phylogenetic analysis, using a bootstrap analysis. From a theoretical point of view, one issue to be inquired is the possibility of using more sophisticated fusion techniques, such as those recently introduced in the clustering contexts based on evidence accumulation.[23] Moreover, the theoretical effects of the fusion of information derived from nucleotide and amino acid mutation models have to be investigated, carefully clarifying the phylogenetic implications of the fusion process.

As a final remark, we would like to stress the fact that the proposed approach could also be useful in the phylogenomic context, as it would allow us to maximize the information extracted for each gene before the combination of results of different genes.

Acknowledgments

The authors are very grateful to Prof. V. Murino for helpful discussions. Moreover, the authors want to thank the anonymous reviewers for their really precious suggestions.

References

1. Bull AT, Ward AC, Goodfellow M, Search and discovery strategies for biotechnology: The paradigm shift, *Microbiol Mol Biol Rev* **64**:573–606, 2000.
2. Fox GE, Stackebrandt E, Hespell RB, Gibson J, Maniloff J, Dyer TA, Wolfe RS, Balch WE, Tanner RS, Magrum LJ, Zablen LB, Blakemore R, Gupta R, Bonen L, Lewis BJ, Stahl DA, Luehrsen KR, Chen KN, Woese CR, The phylogeny of prokaryotes, *Science* **209**(4455):457–463, 1980.
3. Prohaska SJ, Fried C, Wagner GP, Stadler PF, Surveying phylogenetic footprints in large gene clusters: Applications to hox cluster duplications, *Mol Phylogenet Evol* **31**(2):581–604, 2004.
4. Zuckerkandl E, Pauling L, Molecules as documents of evolutionary history, *J Theor Biol* **8**:357–366, 1965.
5. Fitch WM, Margoliash E, Construction of phylogenetic trees, *Science* **155**(760):279–284, 1967.
6. Sails AD, Swaminathan B, Fields PI, Utility of multilocus sequence typing as an epidemiological tool for investigation of outbreaks of gastroenteritis caused by *Campylobacter jejuni*, *J Clin Microbiol* **41**:4733–4739, 2003.
7. Tautz D, Arctander P, Minelli A, Thomas R, Vogler AP, DNA points the way ahead in taxonomy, *Nature* **418**(6897):479, 2002.
8. Jukes TH, Cantor CR, *Mammalian Protein Metabolism*, Academic Press, New York, 1969.
9. Kimura M, A simple model for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences, *J Mol Evol* **16**:111–120, 1980.
10. Barry D, Hartigan JA, Statistical analysis of hominoid molecular evolution, *Stat Sci* **2**:191–210, 1987.
11. Kishino H, Hasegawa M, Evolution of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea, *J Mol Evol* **29**:170–179, 1989.
12. Lake JA, Reconstructing evolutionary trees from DNA and protein sequences: Paralinear distances, *Proc Natl Acad Sci USA* **91**:1455–1459, 1994.
13. Galtier N, Gouy M, Inferring phylogenies from DNA sequences of unequal base compositions, *Proc Natl Acad Sci USA* **92**:11317–11321, 1995.
14. Schwarz R, Dayhoff M, Matrices for detecting distant relationships, in Dayhoff M (ed.), *Atlas of Protein Sequences*, National Biomedical Research Foundation, Washington, DC, pp. 353–358, 1979.
15. Jones DT, Taylor WR, Thornton JM, The rapid generation of mutation data matrices from protein sequences, *Comput Appl Biosci* **8**:275–282, 1992.
16. Yang Z, Phylogenetic analysis using parsimony and likelihood methods, *J Mol Evol* **42**:294–307, 1996.
17. Torriani S, Felis GE, Dellaglio F, Differentiation of *Lactobacillus plantarum, L. pentosus* and *L. paraplantarum* by *reca* gene sequence analysis and multiplex PCR assay with *recA* gene-derived primers, *Appl Environ Microbiol* **67**:3450–3454, 2001.
18. Ronquist F, Huelsenbeck JP, MrBayes 3: Bayesian phylogenetic inference under mixed models, *Bioinformatics* **19**:1572–1574, 2003.
19. Ho TK, Hull JJ, Stihari SN, Decision combination in multiple classifier systems, *IEEE Trans Pattern Anal Mach Intell* **16**(1):66–75, 1994.
20. Kittler J, Hatef M, Duin R, Matas J, On combining classifiers, *IEEE Trans Pattern Anal Mach Intell* **20**(3):226–239, 1998.
21. Melnik O, Vardi Y, Zhang C-H, Mixed group ranks: Preference and confidence in classifier combination, *IEEE Trans Pattern Anal Mach Intell* **26**(8):973–981, 2004.

22. Topchy A, Jain AK, Punch W, Clustering ensembles: Models of consensus and weak partitions, *IEEE Trans Pattern Anal Mach Intell* **27**(12):1866–1881, 2005.
23. Fred A, Jain AK, Combining multiple clusterings using evidence accumulation, *IEEE Trans Pattern Anal Mach Intell* **27**(6):835–850, 2005.
24. Ross A, Jain AK, Multimodal biometrics: An overview, *Proc Eur Signal Processing Conf.*, pp. 1221–1224, 2004.
25. Duin RPW, Tax DMJ, Experiments with classifier combining rules, *Proc Workshop on Multiple Classifier Systems*, pp. 16–29, 2000.
26. Tax DMJ, Breukelen MV, Duin RPW, Kittler J, Combining classifiers by averaging or by multiplying?, *Pattern Recognit* **33**:1475–1485, 2000.
27. Kumar A, Wong DCM, Shen HC, Flynn PJ, Personal verification using palmprint and hand geometry biometric, *Proc Int Conf Audio and Video-based Biometric Person Authentication*, pp. 668–678, 2003.
28. Criscuolo A, Berry V, Douzery EJP, Gascuel O, SDM: A fast distance-based approach for (super)tree building in phylogenomics, *Syst Biol* **55**(5):740–755, 2006.
29. Beal M, Jojic N, Attias H, A graphical model for audiovisual object tracking, *IEEE Trans Pattern Anal Mach Intell* **25**(7):828–836, 2003.
30. Fisher III JW, Darrell T, Speaker association with signal-level audiovisual fusion, *IEEE Trans Multimedia* **6**(3):406–413, 2004.
31. McCowan I, Gatica-Perez D, Bengio S, Lathoud G, Barnard M, Zhang D, Automatic analysis of multimodal group actions in meetings, *IEEE Trans Pattern Anal Mach Intell* **27**(3):305–317, 2005.
32. Cristani M, Bicego M, Murino V, Audio-visual event recognition in surveillance video sequences, *IEEE Trans Multimedia* **9**(2):257–267, 2007.
33. Brunelli R, Falavigna D, Person identification using multiple cues, *IEEE Trans Pattern Anal Mach Intell* **12**:955-966, 1995.
34. Ross A, Jain AK, Information fusion in biometrics, *Pattern Recognit Lett* **24**:2115–2125, 2003.
35. Jain AK, Ross A, Multibiometric systems, *Commun ACM* **47**(1):34–40, 2004.
36. Fumera G, Roli F, A theoretical and experimental analysis of linear combiners for multiple classifier systems, *IEEE Trans Pattern Anal Mach Intell* **27**(6):942–956, 2005.
37. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG, The ClustalX windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools, *Nucleic Acids Res* **25**:4876–4882, 1997.
38. Dellaglio F, Felis GE, Taxonomy of lactobacilli and bifidobacteria, in Tannock GW (ed.), *Probiotics and Prebiotics: Scientific Aspects*, Caister Academic Press, Wymondham, UK, pp. 25–50, 2005.
39. Dellaglio F, Torriani S, Felis GE, Reclassification of *Lactobacillus cellobiosus* Rogosa *et al.* 1953 as a later synonym of *Lactobacillus fermentum* Beijerinck 1901, *Int J Syst Evol Microbiol* **54**(3):809–812, 2004.
40. Naser SM, Vancanneyt M, Hoste B, Snauwaert C, Swings J, *Lactobacillus cypricasei* Lawson *et al.* 2001 is a later synonym of *Lactobacillus acidipiscis* Tanasupawat *et al.* 2000, *Int J Syst Evol Microbiol* **56**(7):1681–1683, 2006.
41. Dellaglio F, Vancanneyt M, Endo A, Vandamme P, Felis GE, Castioni A, Fujimoto J, Watanabe K, Okada S, *Lactobacillus durianis* Leisner *et al.* 2002 is a later synonym of *Lactobacillus vaccinostercus* Kozaki and Okada 1983, *Int J Syst Evol Microbiol* **56**(8):1721–1724, 2006.
42. Duin RPW, The combining classifier: To train or not to train?, *Int Conf Pattern Recognit* **2**:765–770, 2002.

Manuele Bicego received his Laurea degree and Ph.D. in Computer Science from the University of Verona, Italy, in 1999 and 2003, respectively. Now, he is a researcher at the University of Sassari, Italy, and member of the Computer Vision Laboratory. His research interests include statistical pattern recognition, electronic noses, neural networks, hidden Markov models, video analysis, and multiclassifier systems. Recently, he has moved his interest to biometrics, mainly to face recognition, investigating 2D, 3D, and video-based face analysis.

Franco Dellaglio has been a Full Professor of General and Food Microbiology at the University of Verona, Italy, since 1994. He has authored a large number of publications in international journals on different topics related to his diversified interests. His research activities have developed mainly in the fields of taxonomy and application of lactic acid bacteria in different applied fields, such as fermented milks, probiotics, cheeses, and wines.

Giovanna E. Felis received her Ph.D. in Industrial Biotechnology in 2004 from the Science and Technology Department, University of Verona, Italy. Her research has been focused on the taxonomy of different bacteria, and on the contribution of phylogenetic analysis to the development of taxonomic frameworks. After a postdoctoral fellowship in the Food Microbiology Laboratory of the University of Verona, she has also collaborated with the University of Sassari, Italy, and with important international companies.